GETTING STARTED WITH ChIP-SEQ

INTRODUCTION TO ChIP SEQ

Chromatin-immunoprecipitation (ChIP) followed by sequencing of the immuno-precipitated DNA is a powerful tool for the investigation of Protein:DNA interactions. To perform ChIP-seq, chromatin is isolated from cells or tissues and fragmented. Antibodies against chromatin associated proteins are used to enrich for specific chromatin fragments. The DNA is recovered, sequenced and aligned to a reference genome to determine specific protein binding loci. ChIP studies have increased our knowledge of transcription factor biology, DNA methylation and histone modifications.

ChIP-seq was first described in 2007 (1). ChIP sequencing (and also microRNA sequencing) was one of the first methods to make use of the power of massively parallel or next-generation sequencing (NGS) to significantly advance real-time PCR and array-based methods. ChIP-seq is a counting assay that uses only short reads to align to the genome, but requires millions of them to provide meaningful data. Fortunately the Solexa 1G NGS gave up to 30M 21-35bp reads per run. Current NGS systems are producing longer and deeper reads e.g. Illumina and Ion Torrent, however most users still produce single-end 35bp reads. Today of course they generate up to 1.5B per run (Illumina HiSeq flowcell, 2012). This advance in technology has allowed projects like the Encyclopaedia of DNA Elements (ENCODE) to generate almost 1250 ChIP-seq datasets (2). The ENCODE consortium also put significant efforts into standardizing experimental procedures. For instance they produced a set of working standards and reporting guidelines designed to test that an antibody is specific to its antigen and has minimal cross-reactivity to other proteins (3.) Other groups have published comprehensive method descriptions that we refer readers to if they would like a more detailed account of the experimental steps (4).

IN THIS GUIDE WE WILL INTRODUCE ChIP SEQUENCING AND OUTLINE KEY STEPS OF THE EXPERIMENTAL PROCESS INCLUDING:

- Experimental design
- Controls for ChIP-seq experiments
- Reference genome alignment of ChIP-seq reads (mapping)
- Background estimation
- Peak finding
- Quality control of ChIP-seq experiments
- Differential binding analysis
- Motif analysis

ChIP-seq may have evolved from microarray analysis but it required a completely new set of analysis tools to make the most of the platform. ChIP-seq analysis begins with mapping of trimmed sequence reads to a reference genome. Next, peaks are found using peak-calling algorithms. To further analyze the data, differential binding or motif analyses are common end points of ChIP-seq workflows. At every stage the choice of method or algorithm and the parameters used affect the downstream results.

Further complicating analysis options is the fact that ChIP-seq experiments can be divided into different classes (5). Some experiments produce clearly defined peaks of a 100-200 base-pairs as typified by transcription factors, e.g. ERα; others produce wider smears of a few to several hundreds of kilobases such as H3K27me3 and lastly those that produce a mix of clearly defined peaks and wider smears, such RNA polymerase II. Most algorithms have been developed for analysis of clearly defined peaks, as these present the opportunity to determine nucleotide resolution of transcription factor binding and motif analysis.

EXPERIMENTAL DESIGN

All experiments should be designed to meet the goals of the study and make best use of the resources available. Novices to ChIP-seq, or investigators that rely on outside sources for sequencing and data analysis, ought to consult with a bioinformaticist to ensure proper experimental parameters and data formats are in place prior to beginning a ChIP-seq project.

Issues to consider are; the sequencing platform to use, although any NGS platform will work, most ChIP-seq users are concerned with generating as many reads as they can as cheaply as possible; the number of sequence reads generated, with 5-10M being considered the minimum and many users generating 20-40M reads as standard; and biological replication, which is important in understanding variation within sample groups and for differential binding analysis to be carried out (6).

CONTROLS FOR ChIP-SEQ EXPERIMENTS

Two types of controls are often used in ChIP-seq studies, primarily because DNA fragmentation by sonication is not a truly random process. An "input" DNA sample is one that has been cross-linked and sonicated but not immuno-precipitated. An IgG "mock"-ChIP uses an antibody that will not bind to nuclear proteins to generate immuno-precipitated DNA that should be random. Because "mock" ChIP's can often produce relatively little amplifiable DNA input controls are more widely used to normalize signal from ChIP enrichment.

REFERENCE GENOME ALIGNMENT OF ChIP-SEQ READS (MAPPING)

The millions of reads generated in each experiment need to be analyzed and that analysis begins with alignment to a reference genome. The SEQanswers SEQwiki, which hosts a table of common tools for ChIP-seq analysis, lists 94 tools with sequence alignment capabilities.

The most widely used for ChIP-seq have been ELAND, MAQ (7) and Bowtie (8). Mapping is generally performed while allowing for a small number (1-3) of sequence mismatches. Different alignment algorithms trade speed for quality of the final alignment, this is partly down to how they use quality values in the sequence data or align to more repetitive regions of the genome.

MAQ makes use of the sequence quality values so that a mismatch at low quality bases is treated differently to a mismatch at high quality bases, assuming that a low quality base-call is more likely a sequencing error.

Bowtie is one of the fastest mapping algorithms. The algorithms also differ in their handling of reads that map to multiple locations, positioning them randomly or arbitrarily. If ChIP-seq experiments are being performed in highly repetitive regions then the use of paired-end sequencing present the opportunity to anchor read-pairs in a non-repeat region of the genome increasing confidence in the final mapping.

BACKGROUND ESTIMATION

ChIP-seq generates sequence from regions specifically, or indirectly, bound to the antibody target (the signal) as well as from background binding of genomic DNA and regions non-specifically bound to the antibody (the noise). Consequently, ChiP-seq libraries need to be sufficiently complex, consisting of billions of unique molecules with distinct 3' and 5' ends. Even high-quality ChIP-seq libraries are likely to consist mainly of noise rather than signal, with 80-90% background signal possible. Peak calling becomes a signal to noise problem. The choice of analysis algorithm and parameters to use also affects the specificity and sensitivity of the experiment. Mapped reads used for downstream analysis can be restricted to those that map to unique genome regions only (high specificity), or reads that are more "promiscuous" mapping to multiple sites in the genome (high sensitivity). Of note, the complexity of the library and noise can be informed by the size of fragments of ChIP'd DNA. Smaller fragments are more readily clonable and therefore complexity increases when chromatin is highly fragmented.

PEAK FINDING

Probably the most discussed issue in ChIP-seq experiments is the best method to find true "peaks" in the data. A peak is a site where multiple reads have mapped and produced a pileup. ChIP sequencing is most often performed with single-end reads, and ChIP fragments are sequenced from their 5' ends only. This creates two distinct peaks; one on each strand with the binding site falling in the middle of these peaks, the distance from the middle of the peaks to the binding site is often referred to as the "shift".

A good understanding of ChIP fragment size helps in locating the specific nucleotide-resolved binding site. This can be done in the wet lab by gel-based methods; alternatively paired-end sequence data allow the fragment size to be calculated directly from the data. This suggests that a mix of primarily single-end reads, with a small percentage of paired-end reads could provide the best data set for analysis.

The large number of different peak-finders is testament to the importance of finding true peaks in ChIP-seq datasets. Choosing the best is almost impossible; a comparison of eleven different peak detecting algorithms did not show one to be overly superior (9). The parameters chosen for peak calling can significantly affect the outcomes so care must be taken that data sets are analyzed using the same methods.

QUALITY CONTROL OF ChIP-SEQ EXPERIMENTS

After sequencing, mapping and peak finding several quality controls can be used to determine if further investigation and ultimately validation of the data are worthwhile. Packages such as FastQC allow raw sequence quality to be assessed. Read count enrichment can be calculated between ChIP and input samples and can help control for biases in the experimental methods. Finally visual inspection of the data allows a simple but effective tool.

DIFFERENTIAL BINDING ANALYSIS

A relatively new technique is the analysis of differential binding that draws much from the analysis of differential gene expression and has similar power to detect biologically meaningful binding changes between samples (10). The DiffBind package allows identification of genomic loci that are differentially bound between two conditions. It was developed around algorithms used for differential gene expression analysis by RNA-seq. These differential methods allow researchers to assess ChIP peaks quantitatively using peak heights. Key to these methods is the normalization of read counts in ChIP-seq datasets and quantile normalization methods similar to those used in microarray analysis are currently employed.

MOTIF ANALYSIS

One of the most common aims of ChIP-seq experiments has been to discover the sequence motifs for protein binding in the genome. The Multiple EM for Motif Elicitation (MEME) algorithm is the most widely adopted tool for motif discovery (11). Often multiple motifs can be found in a single data set and motif analysis can be performed even on low quality ChIP-seq data although the significance of these motifs is likely to be lower.

CHROMATIN STATE

Another useful analysis of ChIP-seq data comes from a systematic approach used by the ENCODE consortium to characterize genomic regions based on histone modification content. (3) Various histone modifications are assayed using modification-specific histone antibodies in ChIP-seq experiments to obtain a profile of that histone mark within a sample. For their own experiments, the consortium has implemented rigorous specificity tests that use arrays of differentially modified histone tail peptides to ensure antibody specificity. They also share common cell sources which are collectively profiled and compared, ensuring consistency between individual experiments. Their current guidelines cover antibody validation, experimental replication, sequencing depth, data and metadata reporting, and data quality assessment. (13) You can access this information through the Human Epigenome Browser at Washington University. (14)

SUMMARY

ChIP-seq has displaced earlier methods to investigate Protein:DNA interactions almost entirely. Being able to analyze these interactions genome-wide has increased our understanding of transcription factor biology, chromatin modification and transcription. This article presents a broad overview of the major issues that need to be considered when designing and executing ChIP-seq experiments. Laboratory methods are now standardized and kits such as EMD Millipore's Magna ChIP-seq chromatin IP and next generation library construction kit make it possible for virtually any lab to perform ChIP and construct an NGS library.

ChIP-seq is a powerful method and is yielding new biological insights (12). Although the focus today is on detecting the more dispersed class of Protein:DNA interactions and on discovery of statistically significant differential binding, analysis methods are still evolving to allow improved analysis. Projects like ENCODE are showing that it is possible to produce very large data sets as long as experiments are carefully controlled, while at the same time developing useful quality control metrics, analysis methods and parameters for the community.

Useful ChIP Seq Antibodies and Reagents

Product Name	Catalog Number
Kits	
Magna ChIP-Seq™ Chromatin Immunoprecipitation and Next Generation Sequencing Library Preparation Kit	17-1010
Magna ChIP™ A/G Chromatin Immunoprecipitation Kit	17-10085
EZ Magna ChIP A/G Chromatin Immunoprecipitation Kit	17-10086
Magna ChIP™ Protein A+G Magnetic Beads	16-663
Magna ChIP™ Protein A Magnetic Beads	16-661
Magna ChIP™ Protein G Magnetic Beads	16-662
ChIP-seq Qualified Antibodies	
ChIPAb+ Trimethyl-Histone H3 (Lys9)	17-625
ChIPAb+ Trimethyl-Histone H3 (Lys4)	17-614
Anti-acetyl-Histone H4 (Lys12) Antibody	07-595
Anti-trimethyl-Histone H3 (Lys4) Antibody	07-473
Anti-acetyl-Histone H4 (Lys8) Antibody	07-328
Anti-trimethyl-Histone H3 (Lys4) Antibody, clone MC315	04-745
Anti-phospho-Histone H3 (Ser10) Antibody, clone MC463	04-817
Anti-acetyl-Histone H3 (Lys14) Antibody	07-353
Anti-dimethyl-Histone H3 (Lys4) Antibody	07-030
Anti-acetyl-Histone H4 (Lys16) Antibody	07-329
Anti-phospho (Ser10)-acetyl (Lys14)-Histone H3 Antibody	07-081
Anti-acetyl-Histone H3 Antibody	06-599
ChIPAb+ EZH2, clone AC22	17-662
Anti-acetyl-Histone H3 (Lys 4) Antibody	07-539
Anti-trimethyl-Histone H3 (Lys9) Antibody, clone 6F12-H4	05-1242
Anti-Myc Tag Antibody, clone 4A6	05-724
Anti-Histone H4 Antibody, pan, clone 62-141-13	05-858
Anti-acetyl-Histone H3 (Lys14) Antibody, clone 13HH3-1A5	MABE351
Anti-CTCF Antibody	07-729
Anti-acetyl-Histone H4 (Lys5) Antibody	07-327
Anti-RNA polymerase II Antibody, clone CTD4H8	05-623
Anti-monomethyl-Histone H3 (Lys27) Antibody	07-448
Anti-E2F-4 Antibody, clone GG22-2A6	05-312
Anti-trimethyl-Histone H3 (Lys27) Antibody	07-449
Anti-dimethyl-Histone H3 (Lys27) Antibody	07-452
Anti-EZH2 Antibody	07-689
Anti-Androgen Receptor Antibody, PG-21	06-680
Anti-Methylcytosine dioxygenase TET1 Antibody	09-872
Anti-phospho-H2A.X (Ser139) Antibody	07-164
Anti-trimethyl-Histone H3 (Lys9) Antibody	07-442
Anti-monomethyl-Histone H3 (Lys4) Antibody	07-436

References

1. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007 Aug; 4(8):651-7)
2. B.Maher. ENCODE: The human encyclopaedia. Nature 2012 Sep 6; 489(7414):46-8)
3. S.G.Landt, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31)
4. D.Schmidt, et al. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods 2009 Jul;48(3):240-8).
5. S.Pepke, B.Wold & A.Mortazavi, Computation for ChIP-seq and RNA-seq studies. Nature Methods: 2009 Nov;6(11 Suppl):S22-32).
6. C.S.Ross-Innes, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature: 2012 481(7381): 389–393).
7. H.Li, J.Ruan, & R.Durbin. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008 Nov;18(11):1851-8)
8. B.Langmead, C.Trapnell & S.L.Salzberg. Ultrafast and memory- efficient alignment of short DNA sequences to the human genome. Genome 2009;10(3):R25).
9. E.G.Wilbanks & M.T.Facciotti. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLoS One. 2010 Jul 8;5(7):e11471)
10. A.F.Bardet A computational pipeline for comparative ChIP-seq analyses. Nature Protocols: 7;1 2012)
11. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization T.L.Bailey & C.Elkan, C. Machine Learning Journal: 21, 51- 83.(1995)
12. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. H.S.Rhee & B.F.Pugh. Cell: 2011 Dec 9;147(6):1408-19).
13. The NIH Roadmap Epigenomics Mapping Consortium. Bernstein, B.E., (2010) Nature Biotechnology 28(10) 1045-48
14. Zhou, X., et al. (2011) Nature Methods 8(12) 989-990

www.millipore.com/epigenetics

EMD MILLIPORE